The Gold Standard for Assessing Creativity

John Baer, Rider University, Lawrence Township, NJ, USA

Sharon S. McKool, Rider University, Lawrence Township, NJ, USA

ABSTRACT

The most widely used creativity assessments are divergent thinking tests, but these and other popular creativity measures have been shown to have little validity. The Consensual Assessment Technique is a powerful tool used by creativity researchers in which panels of expert judges are asked to rate the creativity of creative products such as stories, collages, poems, and other artifacts. The Consensual Assessment Technique is based on the idea that the best measure of the creativity of a work of art, a theory, a research proposal, or any other artifact is the combined assessment of experts in that field. Unlike other measures of creativity, the Consensual Assessment Technique is not based on any particular theory of creativity, which means that its validity (which has been well established empirically) is not dependent upon the validity of any particular theory of creativity. The Consensual Assessment Technique has been deemed the "gold standard" in creativity research and can be very useful in creativity assessment in higher education.

Keywords: *Consensual Assessment Technique, Creative Products, Creativity Research, Divergent Thinking Tests, Gold Standard*

INTRODUCTION

Assessment of creativity presents a unique challenge in higher education. Although there are tools on the market for assessing creativity, most are designed for young children, and all tend either to lack sufficient validity and reliability or to assess only rather trivial aspects of creativity (or, in many cases, both). If creativity is to be assessed in college settings in a meaningful way, divergent-thinking tests like the Torrance Tests of Creative Thinking and other commonly used creativity tests are inadequate because they fail to meet even the loosest standards of validity (Baer, 2008, 2009, 2011a, 2011b). Self-report measures of creativity and global assessments of students' creativity by others (such as teachers) have also failed to demonstrate sufficient validity to be trusted for most uses. (Baer, 1993; Kaufman, Plucker, & Baer, 2008). Despite the importance of creativity, its assessment has proven to be extremely difficult (Baer, 2011c, 2011d).

The Consensual Assessment Technique is a fairly new method of measuring creativity that could open up new avenues for creativity assessment in higher education. First proposed by Teresa Amabile in 1982 and further developed

DOI: 10.4018/ijqaete.2014010104

by her and other researchers in the last quarter century (Amabile, 1982, 1983, 1996; Baer, 1993, 1994a, 1994b; Baer, Kaufman, & Gentile, 2004; Hennessey, 1994; Kaufman, Baer, Cole, & Sexton, 2008; Kaufman, Baer, Cropley, & Reiter-Palmon, in press; Kaufman, Baer, & Skidmore, 2013; Kaufman, Cole, & Baer, 2009), the Consensual Assessment Technique is now a well validated tool for assessing creativity. It has been called the "gold standard" of creativity assessment (Carson, 2006), but its use has been limited primarily to research settings. It can be used in any field; for example, it can be used for judging the creativity of (a) students' research designs or theories in science, (b) their artistic creations and their musical compositions, or (c) the poems, stories, and essays that they write. It therefore has enormous potential for assessing creativity in higher education settings.

BACKGROUND

Why do you believe that Van Gogh's paintings of sunflowers are creative? On what basis do you judge the special theory of relativity to be highly creative? Why do you think Shakespeare was a more creative dramatist than Marlowe? And how would you judge the creativity of some recent ten- and eleven-dimensional string theories?

You may be comfortable answering some of these questions, but unless you are truly a Renaissance person, it's unlikely that you feel qualified to make a defensible response to all four of them. And even though you might know enough about, say, the works of Shakespeare and Marlowe to give an informed opinion, does your opinion really "count" as much as the opinions of recognized experts in the field of English literature?

How is creativity judged at the highest levels? Why are some works of art treasured and others forgotten? Why do some theories, compositions, books, and inventions win prizes? These kinds of decisions aren't based on a procedure or rubric that awards points for different attributes of a painting, composition, or theory. There is no test to determine which historian's theories, which biochemist's models, or which screenwriter's movies are the most creative. Nobel Prize committees don't apply rubrics, complete checklists, or score tests. What do they do? They ask experts. The most valid assessment of the creativity of an idea or creation in any field is the collective judgment of recognized experts in that field. And while it's true that experts in different times and places may come to different conclusions (and pity the unfortunate artists and scientists whose genius is only recognized when it is too late for them to enjoy their posthumous fame), at any given time, the best judgment one can make of the creativity of anyone's ideas, poems, theories, artworks, compositions, or other creations is the overall judgment of experts in their field[1].

The Consensual Assessment Technique is based on the rather simple idea that the best measure of the creativity of a work of art, a theory, or any other artifact is the combined assessment of experts in that field. Whether one is selecting a short story for a prestigious award or judging the creativity of the painting in an undergraduate art show, one doesn't compute a creativity score by following some checklist or applying a general creativity-assessment rubric. The most valid judgments of the creativity of such artifacts that can be produced -- imperfect though these may be -- are the combined opinions of experts in the field. That's what most prize committees do (which is why only the opinions of a few experts matter when choosing, say, the winner of the Fields Medal in mathematics -- the opinions of the rest of us just don't count). The Consensual Assessment Technique uses essentially the same procedure to judge the creativity of more everyday creations.

Creativity assessment is made difficult by many things, not the least of which are disagreements about the nature of creativity. One of the most fundamental questions in creativity theory and research is the issue of domain specificity. Are the skills, talents, personality characteristics, ways of thinking, and other determinants of creative performance *general-purpose* traits

that a person possessing them can bring to bear on any kind of task? Can one's creativity as a composer of music help her produce more creative paintings? Can one's creativity as a chef help him write more creative short stories? Is a creative biologist likely also to be rather creative as a teacher, a poet, and a dancer? Or, on the other hand, is creativity quite *domain specific*, such that whatever leads to creativity in one domain may be different from that which leads to creativity in other domains?

In the only Point-Counterpoint exchange in its history, the *Creativity Research Journal* asked two leading researchers in the field to make the case for these opposing conceptualizations of creativity (Baer, 1998a; Plucker, 1998). This issue remains unresolved (for recent developments, see Baer, 2012, 2013a, 2013b, 2013c, 2013d; Baer & Kaufman, 2005; Kaufman & Baer, 2005a), and because most creativity tests are tied to one or the other of these models (almost all assume domain-generality, which until recent years was the most commonly accepted hypothesis), the validity of creativity assessment is tied to the validity of particular models of creativity (in addition to all the usual issues that validity raises regarding any test).

Unlike just about every other technique for creativity assessment, the Consensual Assessment Technique is *not* tied to any particular theory of creativity[2]. It works equally well no matter how the domain generality/specificity issue may one day be resolved (or not resolved; as in many contentious issues, the truth is probably somewhere in between this polarity, and the most likely resolution is perhaps a hierarchical model of some type that includes both domain-general and domain-specific features, such as the theory proposed by Kaufman and Baer (2005b; see also Baer & Kaufman, 2005)). The Consensual Assessment Technique is based on *actual creative performances or artifacts*, and it mimics the way creativity is assessed in the "real world." This approach is not without limitations, however. The Consensual Assessment Technique relies on comparisons of levels of creativity *within a particular group*, and it is therefore not possible to create any kind of standardized scoring using Consensual Assessment Technique ratings that might allow comparisons to be made across settings. Its widest use to date has been in research, but it can also be used for many kinds of assessment in higher education, as will be explained below.

PROCEDURES FOR USING THE CONSENSUAL ASSESSMENT TECHNIQUE

The basic technique is quite simple:

1. Subjects are asked to create something (e.g., a poem, a short story, a collage, a composition, an experimental design);
2. Experts in the domain in question are then asked to evaluate the creativity of the things they have made.

The experts work independently and do not influence one another's judgments in any way. The most common kinds of tasks have been writing poems, creating collages, and writing short stories, but the potential range of creative products that one could use is quite wide. No attempt is made to measure some skill, attribute, or disposition that is *theoretically linked* to creativity; instead, it is the *actual* creativity of things that subjects have produced that is assessed. The focus is therefore on creative products, not creativity-relevant talents or attributes that are hypothesized to influence creativity. It is the product or performance itself that is of interest. As Csikszentmihalyi (1999) wrote, "If creativity is to have a useful meaning, it must refer to a process that results in an idea or product that is recognized and adopted by others. Originality, freshness of perception, and divergent-thinking ability are all well and good in their own right, as desirable personal traits. But without some sort of public recognition they do not constitute creativity. . . The underlying assumption [in all creativity tests] is that an objective quality called 'creativity' is revealed in the products, and that judges and raters can recognize it" (p. 314). So instead of trying to measure things

that might be associated with creativity or that might be predictive of creativity, the Consensual Assessment Technique goes right to the heart of creativity by looking at the creative (or not-so-creative) products that subjects have produced.

Here's the basic Consensual Assessment Technique procedure: Subjects are given some basic instructions and, where necessary materials, for creating some kind of product. All subjects are given the same materials and instructions. Then a group of experts, each working independently of one another, assesses the creativity of those creations. In one study, for example, "students were given a line drawing of a girl and a boy . . . [and] asked to write an original story in which the boy and the girl played some part" (Baer, 1994a, p. 39). Experts in the area of children's writing were then asked to rate the creativity of the stories on a 1.0-to-5.0 scale. (The range of the scale is a matter of choice, but should have at least three score points so that there can be some diversity of ratings. Typically judges are free to use fractions if they choose -- e.g., a judge might give a creativity rating of 3.5 -- but in practice, few judges actually employ fractions even when the option exists.) The judges are *not* asked to explain or defend their ratings in any way, and it is important that no such instructions be given. Judges are simply instructed to use their expert sense of what is creative in the domain in question to rate the creativity of the products in relation to one another. That is, the ratings can be compared only *within* the pool of artifacts being judged by a particular panel of experts. High or low levels of creativity, as revealed by the Consensual Assessment Technique, refer to differences within the group of artifacts judged, not in comparison to any external standard. Judges are asked to use the full scale (that is, not to rate all the artifacts as 1s or 2s, or all as 4s or 5s. The goal is to get ratings of the *comparative* creativity of the things being judged. For this reason, a poem that might be judged to be highly creativity in one group of rather pedestrian poems might receive a much lower creativity rating if it were included in a group of much more creative poems.

VALIDITY AND RELIABILITY

The Consensual Assessment Technique assesses creativity at all levels -- everyday creativity as well as creativity at the highest levels -- in the same way that creativity is assessed at the genius level, by asking experts in that field. This is the standard against which any other judgment of creativity would be measured. Rather than use a test, a rubric, or some other device to approximate the judgments of experts, the Consensual Assessment Technique goes directly to the most validyardstick, the experts in a given domain. It is of course true that experts don't always agree and expert opinion may change over time, but at any point in time there is no more objective or valid measure of the creativity of a work of art than the collective judgments of artists and art critics, just as there is no more valid measure of the creativity of a scientific theory than the collective opinions of scientists working in that field. And for the more everyday, garden-variety creativity of most creativity research and most creativity assessments in higher education, the fact that fields may experience paradigm shifts over time is of little significance because few if any of the products being judged will be at the cutting edge of a domain.

But do experts agree? Are they of one opinion regarding which poems, collages, theories, etc. are the most and least creative? A very large number of studies have shown that they consistently *do* agree, and to a remarkable degree (especially when judging everyday, garden-variety creativity), although of course they do not agree completely (which is why a *group* of experts, working independently, is needed). Inter-rater reliability using the Consensual Assessment Technique is typically measured using Cronbach's coefficient alpha, the Spearman-Brown prediction formula, or the intraclass correlation method. These methods generally yield similar inter-rater reliability estimates. Amabile (1983) described a series of 21 studies of artistic (collage-making) and verbal (poetry-writing and story-telling) creativity. The inter-rater reliabilities ranged from .72 to .93. In her more recent work Amabile (1996) has found

a similar range of inter-rater reliability correlations (from .70 to .89), and other researchers have generally reported similar inter-rater reliabilities among expert judges, typically in the .70-to-.90 range (e.g., Baer, 1993, 1997, 1998b; Baer, Kaufman, & Gentile, 2004; Conti, Coon, & Amabile, 1996; Hennessey, 1994; Kaufman, et al., 2008; Runco, 1989). Just as longer tests generally have better reliability, the greater the number of judges who assess the products independently, the higher the overall inter-rater reliability correlations. The average number of expert judges reported by Amabile (1966) was just over 10, with a low of 2 and a high of 40.

But perhaps these ratings are really judgments of something other than creativity. To find out, Amabile (1982, 1983) had raters judge creativity and also a number of other attributes of the products they were evaluating. For example, working with the artistic creativity task of collage-making, Amabile found that while experts tended to agree in their judgments of creativity, these creativity ratings were *not* the same as judgments of such attributes as technical goodness (correlation with creativity ratings = .13), neatness (correlation with creativity ratings = -.26), or expression (correlation with creativity ratings = -.05). There were significant positive correlations with many other judgments, such as novel use of materials (correlation with creativity ratings = .81), complexity (correlation with creativity ratings = .76), and aesthetic appeal (correlation with creativity ratings = .43), but these are all aspects of a collage that *should* be related to the creativity of that collage. A factor analysis of 23 different ratings produced two factors, creativity and technical goodness, and a similar study using poetry-writing produced similar results, with three factors emerging: creativity, style, and technical correctness (Amabile, 1983). So the creativity ratings obtained using the Consensual Assessment Technique have been shown to have good discriminant validity and to be assessments of *creativity*, not of unrelated attributes of the artifacts being judged.

Consensual Assessment Technique ratings of stories, collages, poems, and many other artifacts have been shown to be highly valid measures of creativity in their respective domains, but a caution is in order. The Consensual Assessment Technique does *not* claim to provide evidence of more general creativity-relevant abilities, a topic about which there has been much debate (see, e.g.., Amabile, 1983, 1996; Baer, 1993, 1994a, 1996, 1998a, 2010; Conti, Coon, & Amabile, 1996; Plucker, 1998; Plucker & Runco, 1998; Runco, 1987). Some have argued that such general creativity-relevant skills simply do not exist, and therefore there is nothing to measure and any creativity tests that purports to measure such a general skill cannot possibly be valid, which is perhaps why it has been so difficult to produce a valid creativity test of that kind (Baer, 2011d, 2013a, 2013b).

This is to many people a counter-intuitive idea. *Of course* creativity (as a general skill or trait) exists, many will protest: we see it all the time. And there are many people who are creative in many areas, and others who seem to show little creativity in any endeavor. But this is exactly what one would expect if creativity were totally domain specific (that is, if creativity in one domain did not predict creativity in other domains). If creativity were totally domain specific, creativity in different domains would be *un*correlated (not negatively correlated). There would therefore be a normal distribution of creativity in each domain, and these abilities would be essentially randomly distributed across domains, with some people evidencing creativity in many areas, most people exhibiting varying levels of creativity across domains, and some people showing very little creativity in any domain[3].

If creativity were a general trait or set of skills that could be applied in any field (so that the *same* creativity-relevant skills could help a person be a more creative dramatist, a more creative chemist, or a more creative accountant), then one could use one's poetry-writing creativity to be a more creative chef. Feist (2004) commented on the long-standing (but now fading) assumption of domain generality of creativity:

It is a very appealing, and ultimately firmly American, notion that a creative person could be creative in any domain he or she chose. All the person would have to do would be to decide where to apply her or his talents and efforts, practice or train a lot, and voila, you have creative achievement. On this view, talent trumps domain and it really is somewhat arbitrary in which domain the creative achievement is expressed. Indeed, we often refer to people as "creative," not as "a creative artist" or "creative biologist" (p. 57).

Feist (2004) went on to dispute this view, however, arguing "that this is a rather naïve and ultimately false position and that creative talent is in fact domain specific. There are some generalizedmentalstrategiesandheuristicsthat do cut across domains, but creativity and talent are usually not among the domain general skills" (p. 57).

The two competing theories -- domain generality and domain specificity -- make very different predictions regarding actual creative performance. Becausedomaingeneralityargues that the same creativity-relevant skills, traits, anddispositionsinfluencecreativeperformance across domains, domain generality predicts that people who are creative in one domain are likely to be creative in many domains. This has allowedresearcherstotestthosetheories. Here's how one creativity researcher summarized how these predictions should differ:

Domain generality would be supported by high intercorrelations among different creative behaviors and a common set of psychological descriptors for those behaviors, while domain specificity wouldbe supported byrelatively low correlations among different behaviors, and a diverging set of psychological descriptors of those behaviors. (Ivcevic, 2007, p. 272)

This has been tested using Consensual Assessment Technique ratings of creativity in diverse domains, and these in fact show very little domain generality. Correlations of ratings of subjects' creativity in different domains tend to hover near zero, especially if differences attributable to general intelligence is removed (Baer, 1992, 1993, 1998a, 2010; Han, 2003; Kaufman& Baer, 2005a; Runco, 1989). Creativity researchers are not in complete agreement on the question of how much domain generality there may be, and the best bet is probably on a hierarchical model of some kind (with some abilities contributing modestly to creativity across domains, others only to creativity with a given domain, and others only on specific tasks within a domain, such as poetry within the largerdomainofcreativewriting; see, e.g., Baer & Kaufman, 2005; Kaufman & Baer, 2005b).

In research assessing the impact of a wide variety of interventions, training, or experimentalconstraintsoncreativeperformance, Consensual Assessment Technique ratings have been shown to work well. The technique is not tied to any one theory of creativity, and because it is uncommitted (and therefore unbiased) regarding most of the big questions in creativity research, it can be used equally well by researchers on either side of most research questions. Consensual Assessment Technique ratings are also generally quite stable across time (Baer, 1994b), but they nonetheless respond well to real within-subject changes in motivation. For example:

1. Amabile (1996) found in a series of studies that experimental conditions that make extrinsic constraints salient (such as offering rewards for completing a task, or leading subjects to expect that their work would be evaluated) leadstogenerallylowercreative performance;
2. Baer (1997, 1998b) discovered that this decrement in creative performance under conditionsofrewardorexpectedevaluation is much more prominent among girls than boys;
3. Baer (1994a) found that increases in skill based on training were very narrowly

domain-specific. Subjects trained using divergent-thinking exercises aimed in poetry-relevant skills wrote more creative poems, but not more creative short stories, than subjects who had not received such training.

This has made the Consensual Assessment Technique useful in assessing the impact of varying constraints on creative performance.

GENDER, RACE, AND ETHNICITY AND THE CONSENSUAL ASSESSMENT TECHNIQUE

Most intelligence, aptitude, and achievement tests report different mean scores for different races, ethnicities, and sometimes genders. The validity of such assessments has been fiercely debated (see, e.g., Gould, 1981; Halpern, 2000; Herrnstein & Murray, 1994; Jacoby & Glauberman, 1995; Pinker & Spelke, 2005), and we won't enter that contentious arena. Consensual Assessment Technique scores, in contrast, show very *little* evidence of differences based on race/ethnicity. Kaufman, Baer, & Gentile (2004) conducted the largest study of this type. They performed three separate analyses of the creativity ratings of 103 poems, 104 fictional stories, and 103 personal narratives written by Caucasian, African American, Latino/a, and Asian eighth-grade students as a part of a study using student work collected by the National Assessment of Educational Progress. Each poem, story, and narrative was rated for creativity by 10 expert sin those areas. There were no significant African American-Caucasian differences, and no gender differences[4], on any of the writing tasks. The only significant difference on any of the tasks was in poetry, where there were small but statistically significant differences between the Latino/a-Caucasian groups and Latino/a-Asian groups. Latr studies have confirme this (Kaufman, Baer, Agars, & Loomis, 2010).

HOW THE CONSENSUAL ASSESSMENT TECHNIQUE IS USED

The Consensual Assessment Technique has been used in many ways:

1. To compare creative performance under different (intrinsic v. Extrinsic) motivational constraints (e.g., Amabile, 1983, 1996);
2. To measure the impact of teaching different skills and content knowledge on creative performance (e.g., Baer, 1993, 2003);
3. To study how varying motivational constraints influence the creativity of boys and girls differently (e.g., Baer, 1997, 1998b);
4. To look for possible gender and ethnicity differences in creativity (e.g., Kaufman, Baer, & Gentile, 2004);
5. To compare and evaluate domain-general and domain-specific models of creativity (e.g., Baer, 1993; Conti, Coon, & Amabile, 1996; Runco, 1987; Ruscio, Whitney, & Amabile, 1998);
6. To study the relationship between process and product in creativity (e.g., Hennessey, 1994);
7. To look at creativity in cross-cultural settings (e.g., Niu, 2007; Niu & Sternberg, 2001);
8. To investigate the long-term stability of creativity in a given domain (e.g., Baer, 1994a); and
9. To analyze ways that people with different levels of expertise in a domain conceptualize creativity differently (e.g., Kaufman et al, 2008; Kaufman, Gentile, & Baer, 2005).

The Consensual Assessment Technique has also been used to judge the creativity of such diverse tasks as dramatic performance (Myford, 1989), musical compositions (Hickey, 2001), mathematical equations created by children and adolescents (Baer, 1993), captions written to pictures (Sternberg & Lubart, 1995), personal

narratives (Baer, Kaufman, & Gentile, 2002), and mathematical word problems (Baer, 1993).

The standard format for the Consensual Assessment Technique is to have experts judge the creativity of products that have been created under identical conditions (with all subjects receiving the same instructions and time limits), but recent research has shown that the Consensual Assessment Technique also works when the things to be judged have been created under different conditions (Baer, Kaufman, & Gentile, 2004). This makes possible such uses as comparing how different prompts or assignments impact creative performance differently.

One important caution: It can be tempting when using the Consensual Assessment Technique to use less-than-expert judges, because assembling panels of experts can be time-consuming and expensive. Recent research has shown, however, that, in most areas, experts' judgments and those of novices in a domain do not match. Quasi-experts -- people who have experience working in the field but who have not quite reached what what might deem expert-level credentials -- can often be used successfully (Baer, Kaufman, & Riggs, 2009; Kaufman & Baer, 2012; Kaufman, Baer, & Cole, 2009)

USING THE CONSENSUAL ASSESSMENT TECHNIQUE IN HIGHER EDUCATION

The Consensual Assessment Technique is not limited to use in fields most commonly associated with creativity, such as the arts and sciences. As Emerson (1837/1998) reminded us, "There are creative manners, there are creative actions, and creative words; manners, actions, words, that is, indicative of no custom or authority, but springing spontaneous from the mind's own sense of good and fair" (p. 4; and four paragraphs later he adds "creative reading as well as creative writing" to the list). One might use the Consensual Assessment Technique to judge the creativity of just about anything in which one finds imaginative or original work, such as wedding cakes, cartoons, or even the graffiti found on the walls of buildings.

To date, the Consensual Assessment Technique has not been widely used in higher education, except as a research tool. Although its primary use has been in research, it has also sometimes been used in elementary and secondary education to judge student creativity in a particular area (or several areas) for such purposes as admission to a program for gifted and talented students.

Here are a few arenas in which the Consensual Assessment Technique could be used in higher education:

1. **Research on the effectiveness of college majors or programs:** Colleges want to know how well they are succeeding in their various missions (an interest accreditation boards share). Nurturing student creativity is a goal of some college programs, and in those areas the Consensual Assessment Technique could be helpful. For example, in a program in which students produce a portfolio of creative work, samples of students' creations from different years in a program could be taken. A group of experts in that field could be asked to rate the creativity of the various creations (not knowing which students produced which work, or in what academic year the work was produced, of course). If the creativity ratings are higher the longer students are in a program -- a very easily computed statistic -- that is very strong evidence that the program is successfully nurturing student creativity. (One could also ask the expert judges to rate the artifacts on other dimensions as well as creativity, of course);
2. **Selection for admission to competitive programs:** Colleges have long been using an informal Consensual Assessment Technique for selecting students for programs in creative writing, music, art, theater, and other areas. Validation of the Consensual

Assessment Technique supports such selection techniques and can help guide their use. We know that it is important to use multiple judges; for the judges to make their creativity ratings independently; and for the judges to do what are in effect blind reviews -- that is, they should not know anything about the candidate other than the work being judged. (This is why selections of musicians these days are now often done with the candidate playing from behind a screen, so that other student characteristics -- such as appearance, gender, race, etc. -- cannot be factors in the judges' decisions);

3. **Evaluations of students in regular courses:** In many courses creativity is one aspect of students' work that is to be evaluated, and in such cases it is often the most difficult evaluation professors need to make. Professors might find it helpful to ask colleagues who do not know the students to make independent judgments of the creativity of students' work. This is a bit tricky because Consensual Assessment Technique ratings are always, in effect, norm-referenced ratings based on comparisons within the group of creations being judged. As such, a moderately creative work that is part of a group of very uncreative works will earn top ratings, but the same work would receive low ratings in a group of very creative works. Because some classes have higher levels of creativity than others, this could lead to unfair grading-on-the-curve kinds of assessments.

To get around this and to make the creativity ratings more criterion-referenced, one can do what testing companies like the Educational Testing Service do to make sure different versions of tests are of equal difficulty, and what holistic rating systems do to make sure that multiple raters are using the same standards. One needs to include in one's sample of work some items whose creativity has been previously assessed and for which one has a creativity rating that one trusts. Including a handful of such items that one knows show varying levels of creativity allows one to make adjustments for the varying creativity of the works being judged. Rather than base one's ratings on how well the students in the class perform in comparison to each other, one can use these extra, previously vetted works as one's standards. If a student's work receives a creativity rating equal to a work that one knows to be highly creative, then that is the "score" one would use, not how well it did in comparison to others in the class. (Of course, norm-referenced and criterion-referenced scores typically line up rather closely, but this technique avoids the danger of mis-judging a student's creativity because of the varying creativity of the group of students who happen to be in her class.)

One can also use the Consensual Assessment Technique to compare the work of students at the beginning and the end of a course, as discussed above in the *research on the effectiveness of college majors or programs* section if students will have produced several different works during the semester:

4. **Selecting winners of prizes, fellowships, and other honors:** Many colleges already use a procedure similar to that used by major prize committees to select winners of competitions -- that is, by having experts in the domain in question judge submissions. Following the procedures of the Consensual Assessment Technique ensures that this process is conducted in a fair and well validated manner. As noted above under *selection for admission to competitive programs*, it is important to use multiple judges, for the judges to make their creativity ratings independently, and for the judges to make their judgments without knowing whose work is whose among the artifacts being judged. In competitions such as these, in which some of the judges may know some of the candidates, blind review is especially important.

CONCLUSION AND RECOMMENDATIONS

The Consensual Assessment Technique is a powerfultoolforassessingcreativity. Ithasbeen well validated and is used widely in creativity research. Unlike most "tests" of creativity, the Consensual Assessment Technique does not measure skills or traits that are hypothesized to be part of creative thinking or performance. TheConsensualAssessmentTechniqueassesses actual creative performance.

TheConsensualAssessmentTechniquehas manypotentialapplicationsinhighereducation assessment, but it is not without limitations and drawbacks. It is very resource intensive: assembling groups of expert judges is not simple and it may be expensive. And one cannot replace expert judges with novices (such as by having students judge one another's work) unlessthestudentsthemselveshaveahighlevel of expertise. While gifted and highly creative students have been shown to rate creativity in ways very similar to experts, college students in general do not (Baer, Kaufman, & Riggs, 2009; Kaufman & Baer, 2012; Kaufman, Baer, & Cole, 2009; Kaufman, Baer, Cole, & Sexton, 2008; Kaufman, Gentile, & Baer, 2005).

The Consensual Assessment Technique is not linked to any particular theory of creativity, and its validity does not rise or fall with the success or failure of any theory. It has also been shown to be free of gender and race/ethnicity biases. It has great potential for creativity assessment in many areas of higher education.

REFERENCES

Amabile, T. M.(1982). Socialpsychologyofcreativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*, 997–1013. doi:10.1037/0022-3514.43.5.997

Amabile, T. M. (1983). *The social psychologyofcreativity*. New York: Springer-Verlag. doi:10.1007/978-1-4612-5533-8

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Boulder, CO: Westview.

Baer, J. (1993). *Creativity and divergent thinking: A task-specific approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baer, J. (1994a). Divergent thinking is not a general trait: A multi-domain training experiment. *Creativity Research Journal*, *7*, 35–46. doi:10.1080/10400419409534507

Baer, J. (1994b). Performanceassessmentsofcreativity: Dotheyhavelong-termstability? *Roeper Review*, *7*(1), 7–11. doi:10.1080/02783199409553609

Baer, J. (1996). Theeffectsoftask-specificdivergent-thinking training. *The Journal of Creative Behavior*, *30*, 183–187. doi:10.1002/j.2162-6057.1996.tb00767.x

Baer, J. (1997). Gender differences in the effects of anticipated evaluation on creativity. *Creativity Research Journal*, *10*, 25–31. doi:10.1207/s15326934crj1001_3

Baer, J. (1998a). The case for domain specificity in creativity. *Creativity Research Journal*, *11*, 173–177. doi:10.1207/s15326934crj1102_7

Baer, J. (1998b). Gender differences in the effects of extrinsic motivation on creativity. *The Journal of Creative Behavior*, *32*, 18–37. doi:10.1002/j.2162-6057.1998.tb00804.x

Baer, J. (2003). Impact of the core knowledge curriculum on creativity. *Creativity Research Journal*, *15*, 297–300. doi:10.1080/10400419.2003.9651422

Baer, J. (2005, August 18-21). Gender and creativity. In *Proceedingsofthe Annual Meetingofthe American Psychological Association*, Washington, DC.

Baer, J. (2008). Divergent thinking tests have problems, but this is not the solution. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 89–92. doi:10.1037/1931-3896.2.2.89

Baer, J. (2009). Are the torrance tests still relevant in the 21st century? In *Proceedings of the Annual Meeting of the American Psychological Association*, Boston, MA.

Baer, J. (2010). Is creativity domain specific? In J. C. Kaufman, & R. J. Sternberg (Eds.), *Cambridgehandbook of creativity* (pp. 321–341). Cambridge University Press. doi:10.1017/CBO9780511763205.021

Baer, J. (2011a). Four (more) arguments against the torrance tests. *Psychology of Aesthetics, Creativity, and the Arts, 5*, 316–317. doi:10.1037/a0025211

Baer, J. (2011b). How divergent thinking tests mislead us: Are the torrance tests still relevant in the 21st century? *Psychology of Aesthetics, Creativity, and the Arts, 5*, 309–313. doi:10.1037/a0025210

Baer, J. (2011c). Unintentional dogmatism when thinking big: How grand theories and interdisciplinary thinking can sometimes limit our vision. In D. Ambrose, & R. J. Sternberg (Eds.), *How dogmatic beliefs harm creativity and higher-level thinking* (pp. 157–170). New York, NY: Routledge.

Baer, J. (2011d). Why grand theories of creativity distort, distract, and disappoint. *International Journal of Creativity and Problem Solving, 21*(1), 73–100.

Baer, J. (2012). Domain specificity of creativity: Implications for early childhood education. In O. Saracho, & B. Spodek (Eds.), *Contemporary perspectives on research in creativity in early childhood education* (pp. 43–60). Charlotte, NC: Information Age Publishing.

Baer, J. (2013a). Domain specificity and the limits of creativity theory. *The Journal of Creative Behavior, 46*, 16–29. doi:10.1002/jocb.002

Baer, J. (2013b). Domain specificity of creativity: Theory, research, and practice. *Text.* http://www.textjournal.com.au/speciss/issue13/Baer.pdf

Baer, J. (2013c). Teaching for creativity: Domains and divergent thinking, intrinsic motivation and evaluation. In M. Gregerson, H. Snyder, & J. Kaufman (Eds.), *Teaching creatively and teaching creativity* (pp. 175–181). New York, NY: Springer. doi:10.1007/978-1-4614-5185-3_13

Baer, J. (2013d). Thinking critically about creativity: Why domains matter in understanding, assessing, and promoting creativity. In M. Shaughnessy (Ed.), *Critical thinking and higher order thinking: A current perspective* (pp. 117–126). Hauppauge, NY: Nova Publishers.

Baer, J., & Kaufman, J. C. (2005). Bridging generality and specificity: The amusement park theoretical (APT) model of creativity. *Roeper Review, 27*, 158–163. doi:10.1080/02783190509554310

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal, 16*, 113–117. doi:10.1207/s15326934crj1601_11

Baer, J., Kaufman, J. C., & Riggs, M. (2009). Rater-domain interactions in the consensual assessment technique. *International Journal of Creativity and Problem Solving, 19*, 87–92.

Carson, S. (2006). *Creativity and mental illness*. New Haven, CT: Invitational Panel Discussion Hosted by Yale's Mind Matters Consortium.

Conti, R., Coon, H., & Amabile, T. M. (1996). Evidence to support the componential model of creativity: Secondary analyses of three studies. *Creativity Research Journal, 9*, 385–389. doi:10.1207/s15326934crj0904_9

Csikszentmihalyi, M. (1999). Implications of a systems perspective for the study of creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 313–335). Cambridge, UK: Cambridge University Press.

Emerson, R. W. (1837, August 31). "The American scholar." An oration delivered before the Phi Beta Kappa Society, at Cambridge. *Nature; Addresses and Lectures.* Retrieved December 14, 2007, from http://rwe.org/works/Nature_addresses_1_The_American_Scholar.htm.

Feist, G. J. (2004). The evolved fluid specificity of human creative talent. In R. J. Sternberg, E. L. Grigorenko, & J. L. Singer (Eds.), *Creativity: From potential to realization* (pp. 57–82). Washington, DC: American Psychological Association. doi:10.1037/10692-005

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.

Gould, S. J. (1981). *The mismeasure of man*. New York, NY: W. W. Norton.

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Hillsdale, NJ: Erlbaum.

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal, 7*, 193–208. doi:10.1080/10400419409534524

Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York, NY: The Free Press.

Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education, 49*, 234–244. doi:10.2307/3345709

Ivcevic, Z. (2007). Artistic and everyday creativity: An act-frequency approach. *The Journal of Creative Behavior*, *41*, 271–290. doi:10.1002/j.2162-6057.2007.tb01074.x

Jacoby, R., & Glauberman, N. (1995). *The bell curve debate*. New York, NY: Times Books.

Kaufman, J. C., & Baer, J. (Eds.). (2005a). *Creativity across domains: Faces of the muse*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kaufman, J. C., & Baer, J. (2005b). The amusement park theory of creativity. In J. C. Kaufman, & J. Baer (Eds.), *Creativity across domains: Faces of the muse* (pp. 321–328). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative? *The Journal of Creative Behavior*, *24*, 83–91. doi:10.1080/10400419.2012.649237

Kaufman, J. C., Baer, J., Agars, M. D., & Loomis, D. (2010). Creativity stereotypes and the consensual assessment technique. *Creativity Research Journal*, *22*, 200–205. doi:10.1080/10400419.2010.481529

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior*, 223–233. doi:10.1002/j.2162-6057.2009.tb01316.x

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, *20*, 171–178. doi:10.1080/10400410802059929

Kaufman, J. C., Baer, J., Cropley, D., & Reiter-Palmon, R. (in press). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*.

Kaufman, J. C., Baer, J., & Gentile, C. A. (2004). Differences in gender and ethnicity as measured by ratings of three writing tasks. *The Journal of Creative Behavior*, *39*, 56–69. doi:10.1002/j.2162-6057.2004.tb01231.x

Kaufman, J. C., Baer, J., & Skidmore, L. E. (2013). Young and old, novice and expert: How we evaluate creative art can reflect practice or talent. In S. B. Kaufman (Ed.), *The complexity of greatness: Beyond talent or practice* (pp. 71–82). Oxford University Press. doi:10.1093/acprof:oso/9780199794003.003.0005

Kaufman, J. C., Cole, J. C., & Baer, J. (2009). The construct of creativity: Structural model for self-reported creativity ratings. *The Journal of Creative Behavior*, *43*, 119–134. doi:10.1002/j.2162-6057.2009.tb01310.x

Kaufman, J. C., Gentile, C. A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly*, *49*, 260–265. doi:10.1177/001698620504900307

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. New York, NY: Wiley.

Myford, C. M. (1989). *The nature of expertise in aesthetic judgment: Beyond inter-judge agreement*. Unpublished doctoral dissertation, University of Georgia.

Niu, W. (2007). Individual and environmental influence of Chinese creativity. *The Journal of Creative Behavior*, 151–175. doi:10.1002/j.2162-6057.2007.tb01286.x

Niu, W., & Sternberg, R. J. (2001). Cultural influence of artistic creativity and its evaluation. *International Journal of Psychology*, *36*(4), 225–241. doi:10.1080/00207590143000036

Pinker, S., & Spelke, E. (April 22, 2005). *The science of gender and science: Pinker vs. Spelke: A debate sponsored by Harvard's Mind Brain and Behavior Inter-Faculty Initiative.* Retrieved May 11, 2006, from http://www.edge.org/3rd_culture/debate05/debate05_index.html

Plucker, J., & Runco, M. (1998). The death of creativity measurement has been greatly exaggerated: Current issues, recent advances, and future directions in creativity assessment. *Roeper Review*, *21*, 36–39. doi:10.1080/02783199809553924

Plucker, J. A. (1998). Beware of simple conclusions: The case for the content generality of creativity. *Creativity Research Journal*, *11*, 179–182. doi:10.1207/s15326934crj1102_8

Runco, M. A. (1987). The generality of creative performance in gifted and nongifted children. *Gifted Child Quarterly*, *31*, 121–125. doi:10.1177/001698628703100306

Runco, M. A. (1989). The creativity of children's art. *Child Study Journal*, *19*, 177–190.

Ruscio, J., Whitney, D. M., & Amabile, T. M. (1998). Looking inside the fishbowl of creativity: Verbal and behavioral predictors of creative performance. *Creativity Research Journal, 11*, 243–263. doi:10.1207/s15326934crj1103_4

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd*. New York, NY: Free Press.

ENDNOTES

[1] Even within a given field, different experts might be more appropriate for judging different kinds of works. For example, Pulitzer Prize committees might not be ideal judges of the creativity of compositions by 12-year-old writers; it might be better in that case to have writers and critics who also have familiarity with writings by students of that age serve as judges. Similarly, one might find judgments of the Academy of Motion Picture Arts and Sciences or the Directors Guild useful for judging the creativity of a film, but for judging a film's likely commercial success (or its entrepreneurial film-making creativity) one might instead consult the People's Choice Awards.

[2] Tests of divergent thinking -- the most commonly used tools for measuring creativity -- are examples of a kind of creativity test that *is* anchored to a particular theory of creativity. Divergent thinking tests that ask test-takers to do things like list as many uses for empty tin cans as they can in a short period of time. The theory behind these tests claims that (a) this kind of thinking is important in creativity and (b) the particular content or domain from which the exercise is drawn does not matter. If this kind of divergent thinking is an important component of creativity, *and* if it doesn't matter what domain one uses to test it, then divergent thinking tests might indeed be valid measures of creativity. But if either the divergent thinking theory is wrong or the domain generality theory of creativity is wrong, then these tests cannot be valid ways to assess creativity. In contrast, the validity of the Consensual Assessment Technique is not dependent on the validity of any theory of creativity. It is equally valid no matter which creativity theories prove to be most useful or widely accepted, and because it is not linked to any theory, it can also be used to compare and evaluate theories.

[3] This argument is parallel to that made by Gardner's (1983) Theory of Multiple Intelligences. Gardner argues that his intelligences are orthogonal, and therefore one should expect essentially zero correlations between any two intelligences. That does not mean that there will not be some people who have a great deal of all eight intelligences, however (or some who might score low on all eight). It simply means that the intelligences are randomly distributed, and one's level of intelligence in one area does not in any way predict one's levels of intelligence in any other areas. Creativity, it has been argued, shows even more domain specificity than Gardner's eight intelligences (Baer, 1993).

[4] This is in line with hundreds of studies of creativity using a variety of assessment techniques. Gender differences in such studies tend to be the exception, not the rule (Baer, 2005; Baer & Kaufman, 2005).